eRIC: E-Research – Infrastructure and Communication

Manuel Hora

Technical University of Munich, University Library, manuel.hora@ub.tum.de

eRIC: E-RESEARCH – INFRASTRUCTURE AND COMMUNICATION

Manuel Hora
Technical University of Munich, University Library
Arcisstraße 21, 80333 Munich, Germany
manuel.hora@ub.tum.de

Abstract

The eRIC team at the University Library of Technical University of Munich acts as a central hub for questions from our scientists related to research data management. We provide services for all steps of research or forward our customers to specialised units, e.g. for bibliometric analysis, publishing or trainings in information literacy. Our team supports researchers with regard to funding requirements, data management and data publication as well as with project-specific requests.

Two integral platforms for our services are the TUM Workbench for research in progress and the repository mediaTUM, mostly for static data sets and publications. In the paper at hand, we present the features of the TUM Workbench, including (meta-) data organisation, version control, access management, group functions, structuring of tasks, image annotation, data management plans, WebDAV storages and electronic lab notebooks. Subsequently, we will explain the interplay between TUM Workbench and mediaTUM. With respect to the latter, data can be stored with access restricted to selected individuals, with anonymous access for reviewers or open access with DOI or (versioned) Concept DOI. Both systems, TUM Workbench and mediaTUM, are developed by the software engineering department of TUM University Library in collaboration with external software engineers. The latter association allows for a greater scope of scalability, hence sustainability in the use of the pertaining services.

The guiding principle of eRIC is the development of customisable solutions, courses and IT platforms based on our scientists' needs. At the same time, we construct our platforms TUM Workbench and mediaTUM in a generic fashion, facilitating their reuse at other institutions.

Keywords

research data management; virtual research environment; electronic lab notebook; repository; software development

Introduction and background

Typical depictions of the research data life cycle comprise steps as “Plan“, “Collect“, “Quality Assessment“, “Document“, “Use“, “Preserve“, “Reuse“ [Lenhardt, Ahalt, Blanton, Christopherson, & Idaszak, 2014]. Traditional core competences of libraries focus on retrieval and documentation. Although these are crucial aspects of research data management (RDM), they are not sufficient to support scientists during the whole research data life cycle. Most prominently, there is a gap in many libraries‘ service portfolio for aiding researchers during the phase of actual data usage. During this phase, in which researchers invest a lot of time in analysis and processing, they could exercise proper data management. Later on, time pressure for publications and deadlines towards the end of the project as well as fading motivation due to focus on future proposals frequently hamper good data management.

We aim to assist our scientists during all steps of the research data life cycle, especially with data handling during research in progress. To this end, a joint project of librarians from Technical University of Munich (TUM) and Cape Peninsula University of Technology (CPUT) has identified skills that librarians need to support scientific work during all stages [Leiss & Lockhart, 2015]. Starting from there and based on a survey with active researchers, TUM library has developed its service portfolio for RDM.

Currently we offer regular courses on data management as well as tailored courses for specific institutes. For specific questions, we operate a first level support service. For research in progress, we develop the web platform “TUM Workbench”, comprising groupware, electronic lab notebooks (ELN), file management, image annotation, task and project management as well as data management plans (DMP). All those modules support good scientific practice by change tracking, explicit versioning and metadata enrichment. While the TUM Workbench supports active research, the repository “mediaTUM” stores completed projects. Originally, it served as a platform for university publications and papers as green open access publications. In the meantime, mediaTUM became well used for data publications, due to liberal handling of large data sets, possibilities for fine-tuned access control including anonymous access for peer-reviewers and implementation of Concept DOI for versioning.

In the following, we will present our services in more detail, with a focus on our software platforms. Furthermore, we will discuss the associated procedures for realising new services, the collaboration with IT developers in industry and the transferability of our approaches and services to other libraries.

Support during all steps of a research project

Within our university, the research data service centre is prominently advertised as the first contact point for questions on handling of research data. The official responsibility was assigned in the TUM policy for RDM (https://www.it.tum.de/en/projects/research-data-management/). This policy brought us a remarkable increase in visibility. While we collaborate with several other departments both in the library and in other divisions of TUM, it is convenient for researchers to have one central point of contact for their questions.

Our scientists typically contact us via an e-mail to the RDM queue of our support ticket system. The local RDM team handles most requests. Frequent topics are data publications, aid with DMPs, questions on the handling of TUM workbench, requests for custom RDM courses or invitations to present our electronic services. Occasionally, we receive questions regarding grant applications, suggestions for new features or services or special questions on data management. Some wishes are forwarded to other departments, e.g. questions on general information literacy, bibliometric advice or (open access) publishing.

Characteristics of TUM Workbench

As described above, RDM relies on proper documentation of experiments and annotation of data. We are developing the TUM Workbench as a virtual research environment that facilitates handling of files and data on a day-to-day basis. To motivate researchers to actually use RDM software, it needs to provide clear and visible benefits right from the beginning. The TUM

Workbench is designed such that valuable metadata are added without much effort by the researcher, thus maximising the long-term benefits.

Only such an approach, which provides benefits already in the present, can act as a motivation to actually using RDM software during the research phase. The TUM Workbench is then designed in such a manner, that useful metadata are gathered without much effort from the researcher, thus maximising the long-term benefits.

Providing some basic facts, the TUM Workbench is a web service. Thus, it is independent of operating systems and requires only minor efforts for testing in different browsers. The code is written in Python 3 and Django. While the TUM Workbench is still in an early stage of development, we intend to publish the source code under a copyleft license. The software integrates within the IT environment of TUM, e.g. by reusing LDAP data. However, this is accomplished in a generic manner, hence it can be reused also at other institutions. The following paragraphs will highlight several features and modules of TUM Workbench. Quite several of them exist already as distinct programs. However, we believe that combining these services in one platform is more convenient for users than transferring data back and forth. Such an approach also ensures that data remain on trusted servers and there are no gaps in documentation of research projects.

The central element to organise one's work in our platform are so-called projects. Scientists can assign elements of all modules to projects. This allows once for grouping related elements and second for easily setting access privileges. Thus, collaborators, both from this university and external ones, can get reading or writing permissions for a whole bundle of entries. Where this is too coarse, creators of a specific element can grant fine-tuned permissions to individual users.

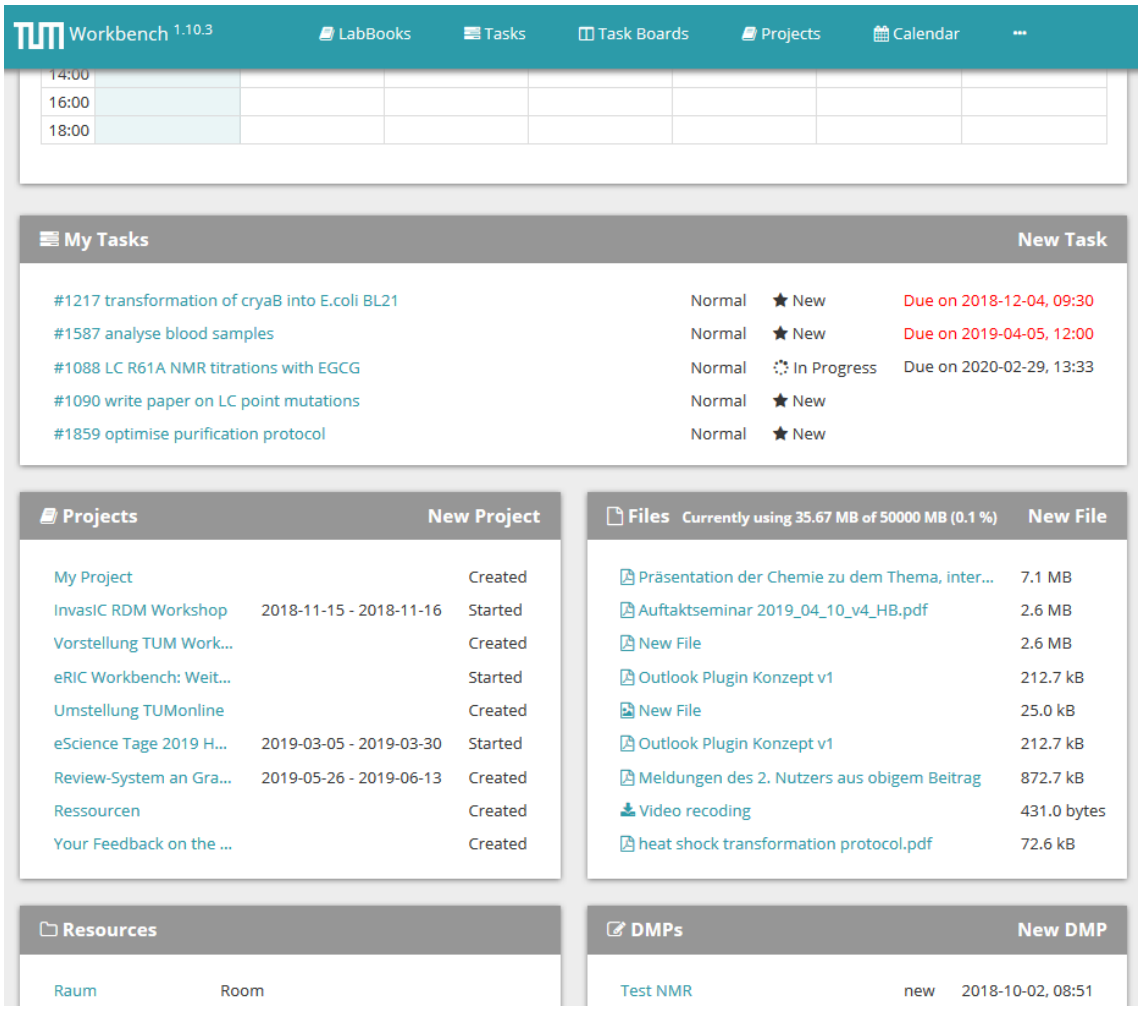

Figure 1: Dashboard of the TUM Workbench

Further modules comprise tasks and Kanban boards, files and storages, meetings and calendar, contacts, DMPs and an ELN. The dashboard (see Figure 1) displays recently used elements from these modules. All instances of these modules have several mandatory metadata fields, often with automatically gathered content. In addition, there are both predefined optional fields and user-defined fields for documentation. Metadata are searchable, also with field-specific options. E.g. it is possible to search for all entries with a numeric metadata field value being lower than 10 AND a certain date field's value being younger than 2019-01-01. Elements have version control and possibilities to highlight changes, where this is reasonable. The latter is especially interesting for collaborative working. All changes are recorded and users can manually create new versions. For those users, who tend to forget about versioning, the software suggests creating a new version after a number of edits to a certain element.

The decision to add features like task management and calendars again follows the notion to create an attractive system, not only for the mandatory documentation, but also to assist with organisation of work. Gathering such information in one system, in turn, makes it easier to recap later. Noteworthy is the function to link elements of all modules with each other. E.g., users can link a document archived in the storages with a meeting, or a task with an entry in the ELN, thereby hinting at yet unresolved questions.

Within the ELN, users can arrange text boxes, images and files on a canvas. For text editing, there is an optional integrated LaTex formula editor with optional GUI. Users can also copy elements from Microsoft Office, including basic mark-up. For images, there is an editor, which supports highlighting, drawing and annotations. While the overall module is denominated as "labbook", its usage is not restricted to laboratory research. Actually, we have received feedback from computer scientists and students of politics, who appreciate this module as a notebook. Many features are useful for a broad variety of disciplines. Figure 2 displays exemplary notes in the ELN. For subject-specific features, we are currently implementing an API for plug-ins. At this point, our close connection to the research departments of TUM is valuable. We are in contact with several of our scientists, who readily agreed to develop plug-ins that tie in with our provided framework. Hence, our approach is to develop generic features and define an API together with the researchers, who contribute discipline-specific plug-ins.

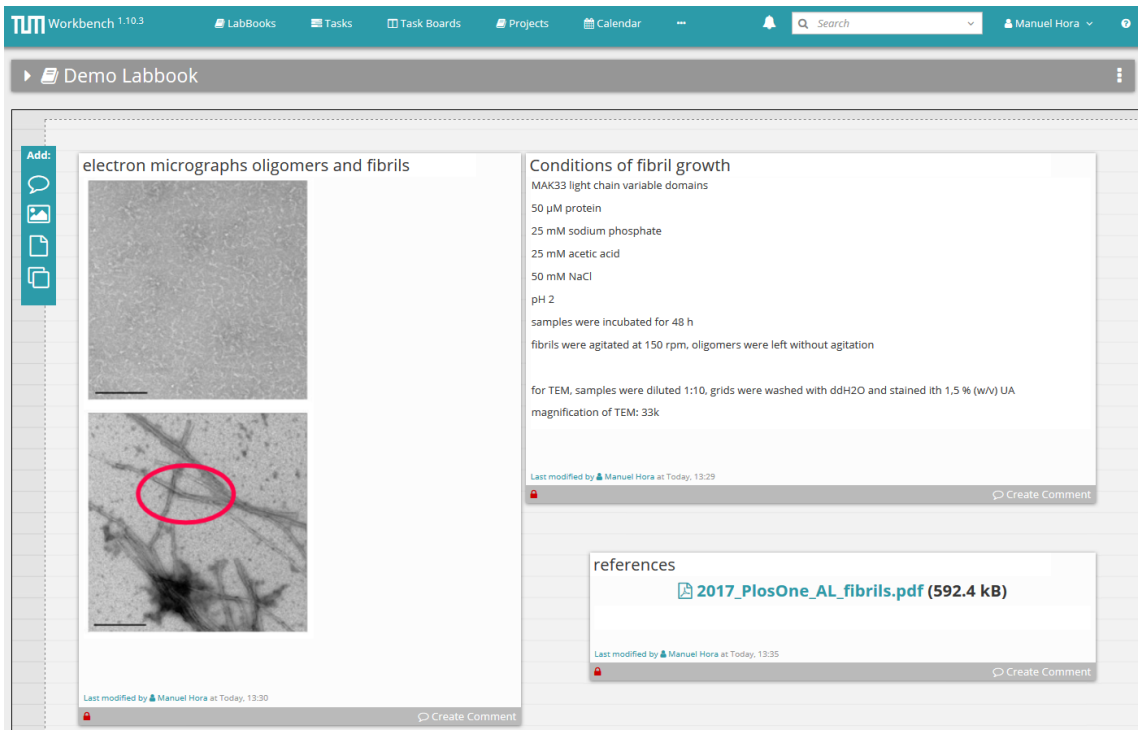

Figure 2: Exemplary entries in the ELN

To summarise, researchers can use the TUM Workbench as a hub for organising and sharing information in a way that enables to retrace the process. The latter option is especially interesting for principal investigators, after PhD students or post docs have left the group.

Features of mediaTUM

Towards the end of a project, quite some scientists publish not only scientific papers, but also the underlying data. For both aims, they can make use of our institutional repository mediaTUM. Our library developed mediaTUM from scratch and thus can readily adapt it to upcoming needs. Interested parties can retrieve the code from GitHub (https://github.com/mediatum/mediatum), published under GPL 3. mediaTUM started as a document repository, same as many others. Later on, collections were added, e.g. an image collection of our architecture museum. Now, we use the repository also for long-term storage and – if desired – publication of data.

We accept large data sets also if a scientist does not intend to publish soon. This is not a technical question but a political one. We have concluded that libraries should offer services for long-term storage independently of publishing decisions. Archiving in a repository ensures proper documentation and enables scientists to grant access later on, if circumstances have changed. In this respect, mediaTUM allows also for some pathways between no access and open to everybody: Access privileges can be assigned to specific research groups or to reviewers, in a manner that maintains anonymity. It is further possible to limit access only to metadata, without granting access to the actual data set.

Since mediaTUM originally handled mostly documents, content was stored on a dedicated drive for the repository. However, this approach reached its limits when users submitted large data sets. Hence, we set up a separate NAS managed by Nextcloud (https://nextcloud.com/files/) for research data. Now, mediaTUM stores metadata and acts as frontend, providing a web link to the data located on the NAS. This allows for convenient storage of data sets comprising several terabytes, a service highly appreciated by scientists who employ high performance computing.

As researchers might want to update or extend data sets after publication, proper referencing benefits from versioned DOIs. Such updates are a frequent issue for scientific software [Smith, Katz, & Niemeyer, 2016], but are not restricted to this field. We implemented the approach known from Zenodo, using Concept DOI [“Zenodo now supports DOI versioning!,” 2017]. In brief, different versions of a data set are assigned individual DOIs, which are all linked to a Concept DOI referring to the data set in general, irrespective of a specific version. Our experience is that scientists are not yet familiar with Concept DOI, but are interested in updating data publications. Although it was implemented in mediaTUM only a couple of months ago, we had already some assignments of Concept DOIs.

Connecting TUM Workbench and mediaTUM

Looking at our current platforms, the obvious next step is to connect TUM Workbench and mediaTUM. Recently, our scientists approached us with a request for tracking data from initial generation up to publication. We are working on such a solution, where data are uploaded automatically from a measurement device to the TUM Workbench, including metadata. Within TUM Workbench, users can track data sets derived from other (primary) data sets. Finally, scientists can see all data related to a given project and decide which of them to transmit to mediaTUM. Thus, user effort for both metadata curation and publishing can be minimised, thereby saving time and removing a major obstacle on the path to more data publications [Stuart et al., 2018].

Development of IT services

While such a connection between a platform for research in progress and a repository can be accomplished for almost any combination of systems, our situation as developers for both platforms allows for deeper integration compared to other constellations. This is one of the benefits of developing, rather than using standard software. Of course, the trade-off are the resources required for creating such platforms. In this regard, we use several approaches here: mediaTUM is mostly developed in-house. In contrast, for the development of TUM Workbench we contracted an external company. A steering committee within the library gathers feature requests and prioritises them. During requirements engineering, we group similar requests and try to anticipate future use cases. Thus, we focus on generic implementations instead of

solutions specific for individual research groups. Once the requirements are defined and approved by both the steering committee and the scientists initially requesting the feature, the task for implementation is forwarded to the developers. With the company, we discuss progress continuously in bi-weekly videoconferences, using a more-or-less agile approach.

Such an assignment to a company allows for fast and scalable development. After only three years, TUM workbench is in productive use by scientists from various disciplines, ranging from biological to political sciences and even non-scientific departments of the university's administration. Similar to mediaTUM, which is already in use at other universities, we seek collaborative use and development of the Workbench with other institutions.

Concluding remarks

This paper has focussed on technical attempts to tackle questions of RDM. We believe that such technical aspects are indeed a crucial component of good RDM. They can save time, which is critical for the economic efficiency of data sharing [Pronk, 2019] and obviously for the acceptance among scientists. In addition, automated technical approaches can increase completeness of data documentation. However, technical efforts cannot be effective on their own. Looking at our RDM services as outlined in the beginning, the workshops for students and scientists are noteworthy with respect to TUM Workbench and mediaTUM. Besides tackling general issues of data management, we also include these platforms in our workshops. Depending on the context, this ranges from awareness raising to advanced questions on the handling. Especially the latter are highly valuable also for us, since such discussions often evolve to new feature requests. Thus, a close connection between RDM trainers and coordinators for RDM software development is beneficial for both services.

Bibliography

Leiss, C., & Lockhart, J. (2015). Librarians' skills for e-research support – joint project at TU München and CPUT. *Proceedings of the IATUL Conferences*, 1–11. Retrieved from http://docs.lib.purdue.edu/iatul/2015/lil/2

Lenhardt, W. C., Ahalt, S., Blanton, B., Christopherson, L., & Idaszak, R. (2014). Data Management Lifecycle and Software Lifecycle Management in the Context of Conducting Science. *Journal of Open Research Software*, *2*(1), e15. https://doi.org/10.5334/jors.ax

Pronk, T. E. (2019). The Time Efficiency Gain in Sharing and Reuse of Research Data. *Data Science Journal*, *18*(2008), 1–8. https://doi.org/10.5334/dsj-2019-010

Smith, A. M., Katz, D. S., & Niemeyer, K. E. (2016). Software citation principles. *PeerJ Computer Science*, *2*, e86. https://doi.org/10.7717/peerj-cs.86

Stuart, D., Baynes, G., Hrynaszkiewicz, I., Allin, K., Penny, D., Lucraft, M., & Astell, M. (2018). *Practical challenges for researchers in data sharing*. https://doi.org/10.6084/m9.figshare.5975011

Zenodo now supports DOI versioning! (2017). Retrieved April 30, 2019, from https://blog.zenodo.org/2017/05/30/doi-versioning-launched/